Title

Automated Detection of Systematic Off-label Drug Use in Free Text of Electronic Medical Records.

Authors

Jung, Kenneth
Lependu, Paea
Shah, Nigam

Publication Date

2013

Peer reviewed

Automated Detection of Systematic Off-label Drug Use In Free Text of Electronic Medical Records

Kenneth Jung, MS[1], Paea LePendu, PhD[1], Nigam Shah, MBBS, PhD[1]
[1]Stanford University, Stanford, CA

Abstract

Off-label use of a drug occurs when it is used in a manner that deviates from its FDA label. Studies estimate that 21% of prescriptions are off-label, with only 27% of those uses supported by evidence of safety and efficacy. We have developed methods to detect population level off-label usage using computationally efficient annotation of free text from clinical notes to generate features encoding empirical information about drug-disease mentions. By including additional features encoding prior knowledge about drugs, diseases, and known usage, we trained a highly accurate predictive model that was used to detect novel candidate off-label usages in a very large clinical corpus. We show that the candidate uses are plausible and can be prioritized for further analysis in terms of safety and efficacy.

Introduction

Off-label use of drugs occurs when a drug is used in a manner deviating from its FDA approved use. Estimates of the extent of off-label use in office-based practices found that 21% of those prescriptions were off-label. Of these usages, 73% lacked adequate evidence regarding safety and/or efficacy (1, 2). Off-label uses are problematic because they have not been evaluated for safety and efficacy. Previous studies relied on surveys of clinicians, had limited coverage in terms of the drugs studied, and have been limited to particular practice types (3).

The widespread adoption of electronic medical records (EMR) provides an opportunity to detect off-label use in an automated, scalable manner. In this paper, we combine features encoding the empirical relationship of mentions of drugs and indications in the free text of clinical notes with additional features that represent prior knowledge about known indications of drugs to build a predictive model achieving high accuracy in a hold out test set. Feature ablation experiments showed that both the empirical features and the prior knowledge derived features were critical to achieving this performance. Notably, our method does not rely on a labeled dataset of clinical text for training the model. We applied this model to a very large clinical dataset to identify potential novel off-label usages. These usages were generally plausible, with some apparently bona fide off-label usages.

Background

Off-label usage of drugs is problematic because such usages have not been evaluated for safety and efficacy. For instance, Tiagabine was approved for use as adjunctive therapy for partial epilepsies. However, when used as the sole or primary treatment, it was found to *cause* seizures. In 1998, 20% of uses of Tiagabine were off-label, but by 2004 this fraction had increased to 94% (4).

Electronic medical records provide an opportunity to detect off-label usage in a comprehensive, automated manner. Unfortunately, EMR systems typically do not link drugs to the indications for which they are prescribed (3). Furthermore, research has shown that the structured data in EMRs is often incomplete, with the free text of clinical notes providing the most complete view of patient care (5). There has been much work done applying Natural Language Processing (NLP) to clinical text for document retrieval and information extraction (6). The 2010 i2b2 NLP Challenge (7) focused on three problems relevant to detecting off-label use — concept recognition; assertion classification; and relationship classification, including the relationship 'Drug used to treat Indication.' If we solved this problem, we could detect off-label usages by simply checking whether these *used to treat* relationships are approved usages. But despite the impressively high performance achieved by submissions to the challenge, these approaches cannot be employed to comprehensively detect off-label usages because they require abundant training data that adequately covers the space of drugs and indications over which we wish to make predictions (8).

In this work, we reframe the problem of detecting off-label drug use to bypass the need for labeled training data. Rather than detecting whether or not a drug is being used to treat an indication within a chunk of text, as in the i2b2 NLP Challenge, we determine whether the drug is being used to treat the indication in the population as a whole. We used a computationally efficient concept extraction pipeline based on the NCBO Annotator (9) Web Service to tag a very large corpus of clinical text from the Stanford Hospital System with mentions of drugs and indications. The empirical counts of mentions from this pipeline have been used for population level tasks such as associating drugs with adverse events — e.g., the relationship between Vioxx and myocardial infarction (10). In particular, these tags have been used to calculate a measure of association between drugs and indications to yield a list of potential off-label usages after applying various heuristic filters (11). We built on this work by combining features

representing the empirical relationship between drug and indication mentions in the text with features encoding knowledge about drugs, indications, and known usage to train a classifier that achieves high performance in a hold out test set. Finally, we applied the classifier to discover novel potential off-label usages.

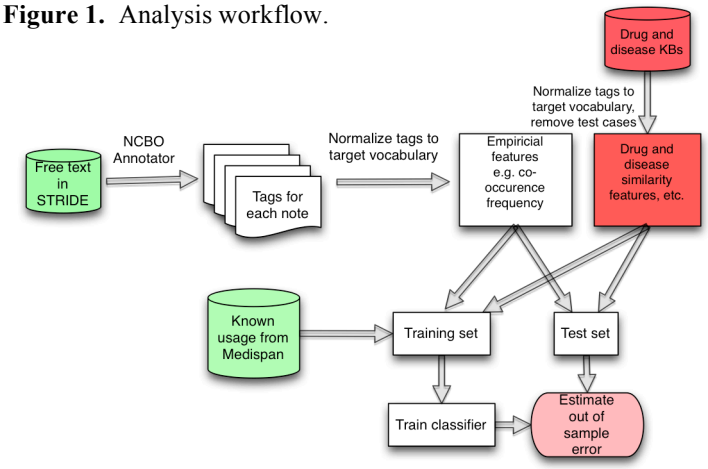
Methods

The learning problem we posed is: predict whether a given drug is being used systematically to treat a given indication. This is a different problem than detecting whether a given drug is being used to treat an indication in a particular clinical note. We are not interested in textual-level drug usage per se, but rather *population level* usage. To this end, we trained a classifier to recognize this population level relationship. Note that we do not predict off-label usage directly - if a drug is being used to treat an indication, we simply look up whether or not the drug is approved for that indication. All other usages are by definition off-label. The overall workflow for our method is depicted in Figure 1 and described in detail below.

Constructing a labeled dataset of positive and negative examples

We used known usages from the Medi-Span® Drug Indications Database™ (Wolters Kluwer Health, Indianapolis, IN) as positive examples. Medi-Span links drugs to indications with one of three relationships: FDA approved, accepted use, and limited evidence. We consider all of these to represent known uses. To construct negative examples, we randomly selected a drug and indication from Medi-Span and then randomly selected another drug and indication that occurs in the data with approximately the same frequency. Frequency based matching was performed because previous work (12) suggests that frequency based features derived from the annotation pipeline can distinguish between drug associated adverse events and treatment relationships. The "negative" pairs were filtered to remove inadvertent known usages. The overall ratio of negative to positive examples in this 'ground truth' dataset was set to 4:1. Negative examples may not be true negatives if they are simply not known by Medi-Span. However, we expect that such pairs are rare so that our labels are a good approximation of the ground truth.

Figure 1. Analysis workflow.



Annotation of Clinical Text

We used an annotation pipeline based on the NCBO Annotator to tag input text with UMLS (13) unique concept identifiers (CUI's), yielding a set of drug and indication mentions for each clinical note. These tags were filtered using NegEx (14) and ConText (15) to remove mentions that are negated or refer to people other than the patient in question. Drugs were normalized into active ingredients (e.g., Excedrin was rewritten into acetaminophen, aspirin and caffeine) using RxNORM (16). Indications were normalized to the set of indications used in Medi-Span using the SNOMED CT hierarchy by recursively rewriting the indication as its parents in the SNOMED CT hierarchy until we reach an indication used by Medi-Span. For instance, *Amok* is not in the Medi-Span target vocabulary so it is rewritten as its parent term, *mania*. We used this pipeline on the free text of clinical notes in the Stanford Clinical Data Warehouse (STRIDE). STRIDE contains over 9.5 million clinical notes for 1.6 million patients over a time period of 18 years. The annotator is flexible with respect to target vocabulary and optimized to run efficiently on very large corpora - this dataset was processed in approximately 7 hours using only 4.5 GB of disk space. We found 1,726 drugs and 1,468 indications that occur in the data.

Feature Construction

For each patient, a drug or indication occurs in their record if it appears at least once in any of the patient's notes; they *co-occur* if they are mentioned in notes that are time-stamped within one day of each other. These counts, along with association measures derived from them — the chi squared statistic, reporting odds ratio and the conditional probability of a drug mention given an indication mention, were used as features. The fraction of patients in which the drug occurs before the mention of the indication (drug first fraction) was also included, along with drug first fractions adjusted for the frequency of the drugs and indications (12). Overall, we calculated nine features that encode the empirical relationship between mentions of the drugs and indications.

We also used features that encode prior knowledge of the drugs, indications and known usage. These features are

motivated by the intuition that drugs are typically used off-label because of similarity with an approved drug, such as a shared molecular target, pathway or drug class — e.g., Bevacizumab, an anti-angiogenic agent, is used off-label to treat age related macular degeneration, a form of blindness caused by aberrant growth of blood vessels (17). We used the Medi-Span and DrugBank databases to construct features that encode this knowledge for each drug-indication pair. For Medi-Span, these included the number of drugs approved for the indication, number of drugs known to be used for the indication, the fraction of known treatments for the indication that are also approved, the maximum similarity of the drug to other drugs known to be used to treat the indication, and the maximum similarity of the indication to other indications treated by the drug. Similarity features were calculated using the cosine and Jaccard similarities as shown in Figure 2 below.

The DrugBank 3.0 (18) database provides information on 6,711 drugs and their molecular targets, pathways, and indications. The annotator was used to map DrugBank drug names and indications to our target vocabulary. Molecular targets, pathways, and drug categories were also extracted for each drug. We calculated similarity features analogous to the Medi-Span similarity features, along with other features that capture similarity with respect to molecular targets, pathways, and drug categories. Note that when calculating these features, we censored the test drug-indication pairs from Medi-Span and DrugBank to prevent data snooping.

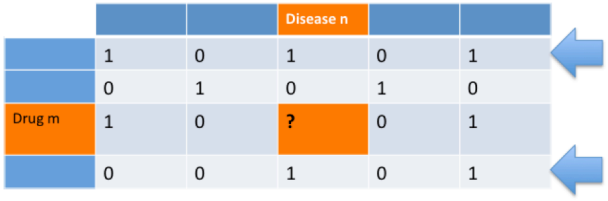


Figure 2. Calculation of drug and disease similarity features. To calculate the similarity of drug m to other drugs used to treat indication n, we find other rows j such that entry (j,n) = 1 (i.e. drug j is known to treat indication n). We calculate the cosine and Jaccard similarities of these rows (indicated by blue arrows) to row m and use the maximum similarity. An analogous calculation is used to calculate indication.

Training a predictive model

We used these 24 features to train a support vector machine (SVM) using the e1071 library in R. We randomly partitioned our 'ground truth' dataset into 37,776 training and 9,446 test examples. An SVM using a radial basis function kernel was trained on the training set and evaluated on the test set. We assessed the contribution of the different features by training and testing models using subsets of the features. We then trained a classifier on the entire labeled dataset. The resulting model was applied to all 2,533,768 possible drug-indication pairs. Predicted usages that are not known usage are assumed to be potential novel off-label usage, and were sorted by estimates of their respective class membership probabilities obtained via a logistic regression model (19). Finally, we normalized indications to remove drug-indication pairs in which the indication is a sub-type of an approved indication.

Results

Table 1 Performance of the classifier on the test set using different subsets of features.

Feature Set	Positive Predictive Value	Specificity	Sensitivity	F1
STRIDE only	0.792	0.969	0.485	0.602
Medi-Span only	0.836	0.988	0.289	0.430
STRIDE + Medi-Span	0.946	0.989	0.768	0.848
STRIDE+DrugBank	0.853	0.977	0.552	0.670
All	0.945	0.989	0.778	0.853

Performance of the classifier

The SVM achieved a precision of 0.945, specificity of 0.989, recall of 0.778 and F1 score of 0.853 on the hold out test set. Feature ablation experiments showed that the different groups of features each contributed significantly to overall performance, particularly with respect to sensitivity (Table 1). For instance, using the empirical (STRIDE) features alone yielded a recall of 0.485, and using Medi-Span features alone yielded a recall of 0.289, while using both feature sets achieved a recall of 0.768. Note that the addition of DrugBank to STRIDE features resulted in significantly better performance than either alone, but DrugBank appears less to be informative than Medi-Span. We speculate that this is because of the relative lack of coverage of DrugBank with respect to indications – DrugBank does not link its indications to SNOMED CT, so we must map them to our target vocabulary using the NCBO Annotator.

Predicting novel off-label usages

We then applied an SVM trained on all labeled examples to predict potential novel off-label usage. This recovered 95.1% of the known uses, and classified 57,749 drug-indication pairs as off-label usages. Applying a probability estimate cut-off of 0.95 and filtering out known usages yielded 10,765 potential novel drug-indication pairs. Manual inspection of the list of potential novel drug-indication pairs revealed that the predictions are generally plausible given prior knowledge of the drugs and indications, and fairly often represent off-label usages that are not in Medi-

Span (see Table 2 below for examples of potential new usages). For instance, (Levofloxacin, Tularemia), (Ciclopirox, Onychomycosis), (Vincristine, Osteosarcoma), and (Cilastatin, Arthritis/Infectious) are predicted off-label usages that are not in Medi-Span and but are supported by PubMed. Cilastatin and 'Arthritis, Infectious' is particularly interesting because Cilastatin is used with the antibiotic, Imipenem, to slow the metabolism of the latter, leading to higher effective concentrations of the drug. This is an example of a drug-indication association that is not precisely a 'used to treat' relationship, but is nevertheless clinically relevant.

Table 2 Examples of predicted novel off-label drug-indication pairs

Drug	***Indication***
Tobramycin	Conjunctivitis
Felbamate	Epilepsies, myoclonic
Calcitriol	Kidney failure
Corticotropin	Brain neoplasms
Gemfibrozil	Hypercholesterolemia
Meclofenamate	Fever
Levofloxacin	Tularemia
Ciclopirox	Onychomycosis
Vincristine	Osteosarcoma
Cilastatin	Arthritis,infectious

Error analysis

Examination of these new pairs reveals several recurring error modes. First, some errors may stem from the annotation pipeline incorrectly identifying concepts. For example, the use of Corticotropin to treat 'Brain neoplasms' may result from tagging 'Corticotropin Releasing Hormone', used to treat swelling secondary to brain tumors, as 'Corticotropin'. Other errors arise from a mismatch between the mapping of indication mentions to their parents and the known indications in Medi-Span (20). For instance, Medi-Span lists Tobramycin as approved for 'eye infections, bacterial', which is not a parent of 'Conjunctivitis', so the latter is considered a novel off-label use. Similarly, Felbamate is approved to treat 'epilepsies, partial' and 'epilepsies, generalized', which are not parents of 'Epilepsies, myoclonic'. A third error mode occurs when a drug is used to treat indications that have sequelae that are also indications, a case of *protopathic bias (21)* - the NSAID Meclofenamate is approved to treat indications that cause fever, which is detected as a novel off-label use.

Discussion

Previous comprehensive surveys of off-label drug usage were derived from the National Disease and Therapeutic Index (IMS Health, Plymouth Meeting, PA), which relies upon surveys of office-based physicians. We have described work that bypasses this expensive, time-consuming method in favor of automated methods that can be applied to the free text in EMRs. Notably, we detect the use of a drug to treat an indication at a population level instead of the textual level, as has been the focus of previous work applying NLP to clinical text. This allows us to use a computationally efficient concept extraction pipeline on a very large corpus of clinical text without labeled training text. Mentions of drugs and indications were augmented with features derived from prior knowledge of known usages, increasing the performance of the classifier over that achieved using only empirical features.

Future work will focus on three areas. First, co-mentions of drugs and indications can be constrained to occur within the same sentence, potentially increasing the specificity of empirically derived features – intuitively, it is likely that a drug used to treat an indication co-occurs in the same sentence. Second, error modes arising from the questionable mappings of known indications to our target vocabulary may be addressed by training a classifier to detect such cases. This could have the additional benefit of flagging regions of disease and drug-indication ontologies that are mismatched in a data-driven manner. Reconciling or combining other sources of known usage, such as the National Drug File - Reference Terminology (NDFRT) (22), which has a well curated set of mappings from drugs to their known indications, may also reduce the false positive rate by improving the filtering step independently of classifier performance. Finally, we ultimately wish to detect novel off-label usage in order to assess the safety and efficacy of such usages. This task would benefit from a systematic method for prioritizing novel off-label usages by criteria that capture their potential risks (23). Intuitively, we are most interested in usages that involve drugs with serious side effects, are very common, or occur in populations with a high degree of poly-pharmacy or co-morbidity, increasing the risk of drug-drug interactions or other uncharacterized adverse events.

Conclusion

Detecting and monitoring off-label use of drugs is an important problem because such uses have not been assessed for safety and efficacy. We have focused on detecting off-label use defined as the use of drugs for unapproved indications. Importantly, we defined the problem as identifying drug-indication pairs at the population level instead of the textual level. A classifier using features that encode the empirical relationship between mentions of drugs and indications in clinical notes, along with prior knowledge about drugs, indications and known usages, achieves high performance in a hold out test set; and yields novel off-label uses that are generally plausible and in many cases appear to be bona fide novel off-label uses.

References

1. Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. Archives of internal medicine. 2006;166(9):1021-6. Epub 2006/05/10.
2. Chen DT, Wynia MK, Moloney RM, Alexander GC. U.S. physician knowledge of the FDA-approved indications and evidence base for commonly prescribed drugs: results of a national survey. Pharmacoepidemiology and drug safety. 2009;18(11):1094-100. Epub 2009/08/22.
3. Dal Pan GJ. Monitoring the safety of medicines used off-label. Clinical pharmacology and therapeutics. 2012;91(5):787-95. Epub 2012/04/05.
4. Flowers CM, Racoosin JA, Kortepeter C. Seizure activity and off-label use of tiagabine. The New England journal of medicine. 2006;354(7):773-4. Epub 2006/02/17.
5. Poissant L, Taylor L, Huang A, Tamblyn R. Assessing the accuracy of an inter-institutional automated patient-specific health problem list. BMC medical informatics and decision making. 2010;10:10. Epub 2010/02/25.
6. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):539. Epub 2011/08/19.
7. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):552-6. Epub 2011/06/21.
8. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):540-3. Epub 2011/08/19.
9. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC bioinformatics. 2009;10 Suppl 9:S14. Epub 2009/09/26.
10. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. Journal of biomedical semantics. 2012;3 Suppl 1:S5. Epub 2012/05/01.
11. Lependu P, Liu Y, Iyer S, Udell MR, Shah NH. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science. 2012;2012:63-70. Epub 2012/07/11.
12. Liu Y, Lependu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science. 2012;2012:47-56. Epub 2012/07/11.
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(Database issue):D267-70. Epub 2003/12/19.
14. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301-10. Epub 2002/07/19.
15. Chapman WW, Chu D, Downing JN. ConText: an algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP. 2007:81-8.
16. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association : JAMIA. 2011;18(4):441-8. Epub 2011/04/26.
17. Martin DF, Maguire MG, Ying GS, Grunwald JE, Fine SL, Jaffe GJ. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. The New England journal of medicine. 2011;364(20):1897-908. Epub 2011/04/30.
18. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research. 2011;39(Database issue):D1035-41. Epub 2010/11/10.
19. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances In Large Margin Classifiers. 1999:61-74.
20. Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. Journal of the American Medical Informatics Association : JAMIA. 2011;18(4):376-86. Epub 2011/05/21.
21. Horwitz RI, Feinstein AR. The problem of "protopathic bias" in case-control studies. The American journal of medicine. 1980;68(2):255-8. Epub 1980/02/01.
22. Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, et al. VA National Drug File Reference Terminology: a cross-institutional content coverage study. Studies in health technology and informatics. 2004;107(Pt 1):477-81. Epub 2004/09/14.
23. Meltzer DO, Hoomans T, Chung JW, Basu A. Minimal modeling approaches to value of information analysis for health research. Medical decision making : an international journal of the Society for Medical Decision Making. 2011;31(6):E1-E22. Epub 2011/06/30.